基于 BPSO 随机子空间的文本情感分类研究

张庆庆 [1,2]　刘西林 [2]
[1](西安工程大学管理学院　西安 710048)
[2](西北工业大学管理学院　西安 710129)

摘要:【目的】针对基于机器学习的文本情感分类研究中的文本特征表示向量高维性问题, 提出 BPSO 与随机子空间方法结合的选择性集成算法。【方法】在分析 BPSO 与随机子空间原理的基础上给出 BPSO 随机子空间的模型框架及算法流程。将中文评论语料进行特征化表示后, 使用 BPSO 随机子空间进行实验验证和分析。【结果】通过改变随机子空间中子空间率的取值, 研究标准随机子空间与 BPSO 随机子空间选择性集成对分类准确率和系统差异度的影响, 结果表明 BPSO 随机子空间无论在分类准确率还是在系统差异度上均高于标准随机子空间。【局限】尚未在英文数据上进行验证。【结论】将 BPSO 应用于随机子空间方法构成一种新颖的选择性集成模型, 不仅解决了特征向量空间高维性的问题, 而且提高了分类的准确率和泛化能力, 为中文文本情感分类提供了有效的方法。

关键词: 随机子空间　BPSO　文本情感分类　子空间率
分类号: TP391.1

1　引　言

互联网提供给人们丰富的信息资源, 其中表达看法、意见、建议等的主观性文本(如科技评论、产品评论、体育评论、时事评论、博客、影视评论、新闻评论、军事评论、音乐评论、股票评论等)是比重较大且十分重要的部分。这些主观信息是针对特定对象而发表的观点、态度、意见、立场等, 有强烈的个人情感色彩。文本情感分类是针对主观性文本进行自动分析、处理归纳的技术。文本情感分类技术在电子商务、电子政务、信息预测等领域有重要的应用价值。

目前用于文本情感分类研究的主流方法是机器学习, 主要针对文本情感分类任务的特征表示和分类模型的应用与改进上。特征表示旨在找到能最大程度代表句子语义、句法的特征项。目前用于文本特征表示的有一元词(Unigram)、多元词(N-grams)、词性(Part-of-Speech, POS)、词的关系特征、基于规则的特征、结合情感词典的特征和社交网络特征等[1-3]。基于依存语法的依存句法关系特征因其能够很好地表达句子句法和词语间修饰关系, 而被用于文本情感分类的特征表示中, 取得了比常用特征更高的分类精度[4]。在分类模型方面, 传统的分类算法有支持向量机、朴素贝叶斯、最大熵模型、k 近邻、决策树算法等。分类算法用于文本情感分类任务的优劣是一个无定论的问题。集成学习利用"多个分类器决策结果的可信度高于单个分类器决策结果"的思想形成一种新的文本分类模式。利用传统的分类算法作为单个分类器的训练算法, Wang 等比较了三种不同集成学习基分类器的生成方法, 即 Boosting、Bagging 和随机子空间方法, 将三种方法的基分类器分别用朴素贝叶斯、最大熵、决策树、k 近邻和支持向量机分类算法进行训练, 得出集成后的系统分类准确率高于单个分类器的分类准确率的

通讯作者: 张庆庆, ORCID: 0000-0002-5507-466X, E-mail: suiyue2959@163.com。

结论[5]。集成学习用于文本情感分类的研究中，引入Boosting 和 Baggging 方法较多，而随机子空间(Random Subspace)方法较少，针对这一问题，且结合随机子空间方法更适合文本情感分类高维性数据的特点，Wang 等提出基于词性分析基础上的随机子空间情感分类方法，以支持向量机(SVM)作为基学习器，此方法取得了比其他分类器更好的实验结果[6]。

以上研究大多数集中于英文文本，鉴于中文语言表达的复杂性，需要更加能够代表句子语义的特征进行表示。针对中文文本研究较少的问题，并结合目前的基于机器学习的文本情感分类研究现状，本文在依存句法解析表示句子特征的基础上采用随机子空间方法对中文文本进行情感分类研究。并在标准的随机子空间方法上采用选择性集成机制。本文将 BPSO (Binary Particle Swarm Optimization)算法用于随机子空间方法中，提出基于 BPSO 算法的随机子空间方法(缩写为 RS_BPSO)，用于文本情感分类任务中。通过选取不同子空间率研究离散二进制粒子群对随机子空间集成学习准确率和差异度的影响。

2 BPSO 随机子空间方法

2.1 随机子空间

随机子空间(Random Subspace, RS)是集成学习中的一种。集成学习通过构建并结合多个学习器完成学习任务，有时也被称为多分类器系统、基于委员会的学习等。集成学习通过将多个学习器进行结合，常可获得比单一学习器显著优越的泛化性能。集成学习的一般结构是先产生一组“个体学习器”，再用某种策略将其结合起来，产生的个体学习器亦称“基学习器”。较早进行集成学习研究的是 Dasarathy 等[7]，之后集成学习成为机器学习一个重要的研究方向[8-9]。

按照个体学习器生成方式，目前的集成学习方法大致可分为两大类：基于数据划分的方法(Data Partitioning Methods)和基于特征划分的方法(Attribute Partitioning Methods)。其中基于数据划分的方法通过处理训练样本产生多个样本集，主要有 Bagging 和 Boosting 方法。基于特征划分的方法把数据特征划分成子集，用作不同分类器的输入向量，每次使用一个特征子集。主要有 Random Subspace 等。Random Subspace 主要通过随机抽取特征子集构造基学习器，因此对于高维数据有更好的适用性[5]。

随机子空间由学者 Ho 在 1998 年提出，其基本思想来源于随机判别分析[10]。随机子空间将所有特征作为一个大的集合，从中随机选择部分特征形成多个特征子集。文本被表示成多个特征子集代表的特征向量空间模型(Vector Space Model, VSM)。通过分类算法学习形成基分类器。基分类器从不同特征子集只能学习到部分样本信息，通过融合多个分类器以利用所有的样本信息。因此，随机子空间不仅能有效降低特征维数，同时可以结合多个基分类器的优势。原始数据中包含的信息通过多个不同子空间的基分类器融合后得以保持。

随机子空间运用自助法(Bootstrap Method)在所有特征集中随机挑选，形成多个不同的特征子空间。对多个不同的子空间用机器学习分类算法进行训练得到多个基分类器。基分类器的融合方式可以是多数投票法、乘法规则等。本文采用多数投票法作为结果的融合结果。其计算方法为：假设有 T 个基分类器，对输入样本 X 的分类结果分别为 $h_t(x_t)$，x_t 对应第 t 个基分类器的输入，集成系统的分类结果为 $H(X)$。Y 为标签类别集合，当 y 对 Y 中所有值进行逐一取值时，对 y 值与基分类器结果相等的个数进行计数，取计数最多的 y 值为最终的分类结果，即 $H(X)$ 的值，如公式(1)所示[5]。

$$H(X)=argmax_{y\in Y}\sum_{t=1}^{T}1(y=h_t(x_t)) \tag{1}$$

其中，$1(\alpha)$ 为示性函数，如果 α 为真，则 $1(\alpha)=1$；否则 $1(\alpha)=0$。

衡量集成学习系统好坏的两个重要指标是分类准确率和系统差异度。互补且精确的分类器集成得到的系统会更优，如果基分类器结果相似，则系统泛化能力不会得到提高；如果基分类器的结果是多样化的，一个被某分类器错分的样本可以被另外一些分类器正确分类，则可以得到正确结果[11]。

在随机子空间方法中有个重要的参数——子空间维数的选取，一般用选取的特征数目与特征总数的百分比表示，这个比例称为子空间率(Subspace Rate)，用 $k(0<k<1)$ 表示。

假设文本被表示为特征维数为 D 的向量空间。对所有的特征维数 D，运用自助法在所有特征集中随机

挑选，形成 n 个不同的特征子空间。根据子空间率 k，得出特征子集维数为 $k \times D$。

为了增强或者调整系统差异度，通常通过改变子空间维数占总体向量空间维数的比例调节。关于 k 值的选取范围，目前还没有统一的标准。子空间率的选择与系统差异度有密切的联系，子空间率能够影响系统差异度。

2.2　选择性集成

选择性集成是指从一批训练好的基分类器中，选择一部分进行集成。选择性集成的概念由 Zhou 等提出，所有分类器个体全部参加不一定能保证集成泛化能力的提高，同时给出选择性集成好于全部个体参加集成的理论分析[12]。选择性集成实际上是一个全局优化的过程。

全局优化是选择性集成研究的重要方向，选择性集成可以方便地转化为一个组合优化的问题进行研究。粒子群优化算法需要的参数少，执行效率高，收敛速度快，而且粒子群优化算法具备全局搜索能力以及在解决组合优化问题上的优势。粒子群用于集成学习的其他方法如 Bagging、Boosting 方法较多[13-15]，用于随机子空间的较少。

2.3　BPSO 算法

BPSO 算法是在基本粒子群优化算法基础上基于连续空间的一种离散化方法，由美国社会心理学家 Kennedy 和电气工程师 Eberhart 专门针对 0-1 整数规划问题而提出的[16]。

基本粒子群优化算法用公式(2)描述。

$$\begin{cases} v_{ij}(t+1) = v_{ij}(t) + c_1 r_{1j}(t)(p_{ij}(t) - x_{ij}(t)) \\ \qquad\qquad + c_2 r_{2j}(t)(p_{gj}(t) - x_{ij}(t)) \\ x_{ij}(t+1) = x_{ij}(t) + v_{ij}(t+1) \end{cases} \tag{2}$$

其中，$v_{ij}(t+1)$ 和 $x_{ij}(t+1)$ 分别表示第 i 个粒子第 j 维速度在第 $t+1$ 次迭代时的速度和位置。c_1、c_2 为加速常数，通常在[0,2]取值，c_1 调节粒子向个体最优位置移动的步长，c_2 调节粒子向全局最优位置飞行的步长。$r_{1j}(t)$ 和 $r_{2j}(t)$ 是区间[0,1]的随机取值，主要为增加粒子飞行的随机性。p_{ij} 和 p_{gj} 分别表示粒子第 j 维的个体极值和全局极值。而下一次迭代的位置信息 $x_{ij}(t+1)$ 则通过在原有位置上进行速度的转变得来。

而在 BPSO 中，每个位置分量 x_{ij} 取值要么为 1，要么为 0，因此速度分量 v_{ij} 不再表示位置变化的大小，它反映的是 x_{ij} 取 1 的概率。使用速度更新公式时，v_{ij} 取值越大，粒子的位置分量 x_{ij} 越有可能取 1；v_{ij} 取值越小，则 x_{ij} 越趋向于取 0。为了使概率值在[0, 1]之间，BPSO 采用 Logistic 变换对 v_{ij} 进行处理，如公式(3)所示。

$$S(v_{ij}) = \frac{1}{1 + \exp(-v_{ij})} \tag{3}$$

其中，$S(v_{ij})$ 表示位置 x_{ij} 取 1 的概率，粒子改变位置值如公式(4)所示。

$$x_{ij} = \begin{cases} 1 & if\ \ rand(\ ) \leqslant S(v_{ij}) \\ 0 & otherwise \end{cases} \tag{4}$$

其中，$rand(\ )$是一个随机数，从区间[0,1]分布中随机产生。为了避免 $S(v_{ij})$ 接近 0 或 1，参数 $v_{\max}$ 作为最大速度值，限制 v_{ij} 的范围。

3　BPSO 随机子空间算法

BPSO 随机子空间方法在随机子空间训练得到的分类器基础上，对基分类器用 BPSO 算法进行优化选择。基于以上对随机子空间和 BPSO 算法的分析，BPSO 随机子空间方法的算法流程如图 1 所示。

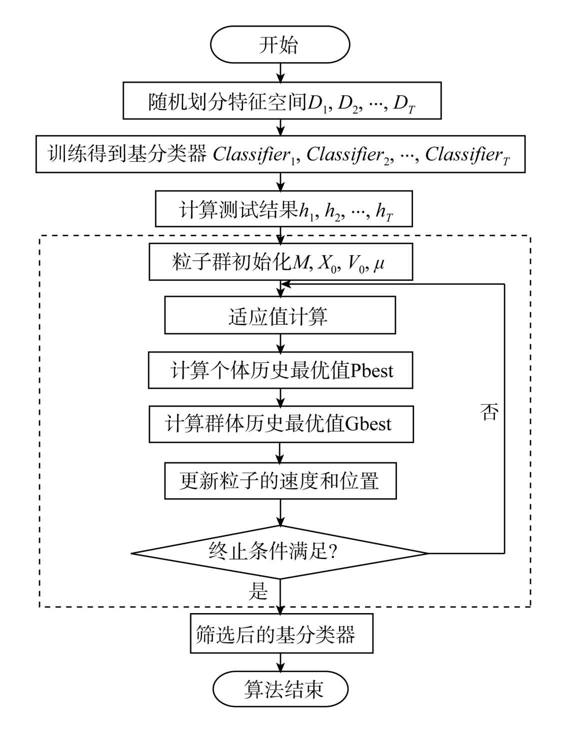


图 1　BPSO 随机子空间算法流程

BPSO 随机子空间方法关键在于 BPSO 中粒子维数、适应值函数的设计。在 BPSO 选择性集成学习算法中, 一个粒子代表对基分类器选择的一种方案。粒子为一个向量, 向量维数则为基分类器的个数, 且与基分类器一一对应。向量值取 0 或者 1。假设基分类器个数为 D 个, 则粒子被表示为一个 D 维向量。如果第 d 个分量的值取 1, 则表示第 d 个基分类器被选中; 反之, 如果第 d 个分量的值取 0, 则表示第 d 个基分类器没被选中。以 10 个基分类器为例, 如果顺序排列, 结果如图 2 所示。

1	2	3	4	5	6	7	8	9	10
0	0	1	1	0	1	0	1	0	1

图 2　基学习器选择结构示意图

10 个基分类器的编号分别为 1 到 10, 经过优化选择后的结果为: [0 0 1 1 0 1 0 1 0 1], 则基分类器被选中的编号分别为 3、4、6、8、10。

在粒子群优化算法中, 粒子的位置信息即为基分类器是否被选中信息, 速度则对应此基分类器被选中的概率。

BPSO 算法根据适应度函数进行全局搜索。集成系统的分类准确率和系统的差异度是判别集成学习系统的两个指标。最常用的评价函数是当前系统的预测误差, 根据系统分类准确率判断集成系统的好坏。另一个评价指标是系统差异度, 系统差异度是衡量集成系统泛化能力的一个指标, 这是一种间接方法, 需要合适的描述才能得到好的结果。本文将采用系统分类结果的准确率作为 BPSO 的适应度函数。

4　实验设计

4.1　数据集

本文数据集来自数据堂提供的情感分析语料[①], 其中包括酒店评论数据、图书评论数据和笔记本电脑评论数据, 分别抓取于携程旅游网、当当网和京东网。

三个原始数据集中的每个数据集均包含 4 000 条正向文本和 4 000 条负向文本, 但都以段落的形式存在。本文研究句子级别的文本情感倾向, 对原有数据进行相应处理。

(1) 对文档数据进行分句处理。以“\n”和中英文的疑问号“？”“?”、句号“。”“.”和分号“；”“;”为断句标识对所有文档进行断句。在断句基础上, 进行去重操作。

(2) 在原有标注文本基础上对断句后的新文本进行甄别, 删除不属于原有类别的句子。

(3) 对三个数据集进行随机抽取。

本文研究平衡数据的文本情感分类, 对酒店评论数据抽取句子 4 000 条, 包括 2 000 条正向评论语句和 2 000 条负向评论语句; 图书评论数据 2 000 条, 包括 1 000 条正向评论数据和 1 000 条负向评论语句; 笔记本电脑评论数据 1 000 条, 包括 500 条正向评论语句和 500 条负向评论语句。

以文献[4]提出的三元组依存关系特征方法为基础, 将中文评论语料转化为三元组依存关系特征。其中, 三个数据得到的特征总数如表 1 所示。

表 1　三元组依存关系特征个数

数据集	三元组依存关系
酒店	140 911
图书	66 297
笔记本电脑	28 932

4.2　评价指标

实验的评价指标采用平均分类准确率和系统差异度。平均分类准确率如公式(5)所示。

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{5}$$

其中, TP(True Positive)表示正向情感文档被正确判断的样本数, TN(True Negative)表示负向情感文档被正确判断的样本数, FP(False Positive)表示正向情感文档被错误判断的样本数, FN(False Negative)表示负向情感文档被错误判断的样本数。TP、TN、FP、FN 的总和是整个待分类文档数。平均分类准确率数值越高, 则对文本的主观性倾向判断越准确。

差异度度量是集成学习系统特有的评估标准, 以下将介绍常用的 4 种成对差异度度量: Q 统计、相关系数 ρ 、不一致度量 dis 和双次失败度量 DF。

假设有 L 个基分类器, C_i 和 $C_j(i, j = 1, 2, \cdots, L,$

①http://www.datatang.com/data/11936.

$i \neq j$) 分别为两个不同的分类器，$N^{11}(N^{00})$ 为分类器 C_i 和 C_j 都对其正确(错误)分类的样例数目；$N^{10}(N^{01})$ 为满足要求的样例数目；分类器 $C_i(C_j)$ 对其正确分类而分类器 $C_j(C_i)$ 对其错误分类。由此，总的样例数目可以表示为 $N = N^{11} + N^{00} + N^{10} + N^{01}$。具体如表 2 中公式(6)–公式(9)所示。

表 2　集成系统差异度度量公式

质量法	公式	编号
Q 统计	$Q_{ij} = \frac{N^{11}N^{00} - N^{10}N^{01}}{N^{11}N^{00} + N^{10}N^{01}}$	(6)
相关系数 ρ	$\rho_{ij} = \frac{N^{11}N^{00} - N^{10}N^{01}}{\sqrt{(N^{11}+N^{10})(N^{01}+N^{00})(N^{11}+N^{01})(N^{10}+N^{00})}}$	(7)
不一致度量 dis	$dis_{ij} = (N^{10} + N^{01}) / N$	(8)
双次失败度量 DF	$DF_{ij} = \frac{N^{00}}{N}$	(9)

由表 2 可以看出，Q 统计数值越大，差异度越低；相关系数 ρ 与 Q 统计有一样的趋势；不一致度量 dis 关注两个分类器分类结果不同的样本，这样的样本越多，差异度越高；而双次失败度量关注两个分类器均将其分类错误的样本，如果这样的样本越多，则准确性和差异度均达到最低。

4.3　实验流程

为了检验基于 RS_BPSO 的有效性，实验将 RS_BPSO 与标准 RS 得到的分类准确率和系统差异度进行比较分析。具体的实验过程如下:

(1) 将评论数据按照 70%和 30%的比例分成训练集和测试集两部分。

(2) 将训练集和测试集的文本转化成结构化的三元组依存关系特征表示的特征向量空间形式。

(3) 将训练集和测试集采用自助抽样法进行特征子集划分。

(4) 对划分过的训练集采用支持向量机进行训练得到基分类器。

(5) 标记所有的基分类器，用 BPSO 算法对基分类器进行优化选择，确定保留的基分类器编号。

(6) 将划分过的测试集用于保留的基分类器中，用多数投票法将得到的基分类器上的测试集的结果进行融合，得到最终的分类结果。

随机子空间方法中，随机选取子特征维数由子空间率 k 决定。取 k 为 0.01、0.02、0.03、0.05 等 4 个值，研究不同 k 取值对文本情感分类准确率和差异度的影响。三个数据集在不同的子空间率取值下，得到的基分类器的特征维数如表 3 所示。

表 3　随机子空间方法下特征子集维数

k	酒店	图书	笔记本电脑
k=0.01	1 409	663	289
k=0.02	2 818	1 326	579
k=0.03	4 227	1 989	868
k=0.05	7 046	3 315	1 447
总个数	140 911	66 297	28 932

表 3 中，可以看出随着 k 的取值不同，不同的数据集特征维数由万维以上降低到千维或者百维。以酒店评论数据为例，当 k 值取 0.01 时，基分类器中用于训练的特征向量空间维数为 1 409。为了所有特征项都有被选取到的可能，对原有特征集进行 50 次随机采样，即特征项被划分为 50 个特征子集，那么训练得到的基分类器个数也为 50。

粒子群优化算法中，适应值函数采用系统分类准确率。为了考察群中粒子个数对文本分类准确率和差异度的影响，设置粒子个数分别为 10、20、30、40、50、60、70、80、90、100 进行比较，迭代次数设为 100，学习因子 $c_1 = c_2 = 2$，惯性权重采用线性迭代 $w_min = 0.1$，$w_max = 0.6$。由于粒子群初始值的随机性，实验结果取 20 次平均值作为最终结果。

5　实验结果及分析

5.1　分类准确率结果及分析

对不同子空间率 k 取值下的标准 RS 和 RS_BPSO 下按照实验流程进行计算，三个数据集的实验结果分

别如表 4–表 6 所示。其中，RS_BPSO 的分类准确率结果一列中括号里标注了优化选择后的平均基分类器个数。

表 4　酒店评论数据分类准确率比较

k	RS	RS_BPSO
0.01	0.6825	0.8342(17)
0.02	0.7183	0.8013(14)
0.03	0.7717	0.8293(13)
0.05	0.8075	**0.8429**(19)

表 4 是酒店评论数据分类准确率结果。可以看出，RS_BPSO 得到的分类准确率均比标准随机子空间方法高，且高出幅度达到 3%-15%。最高分类准确率达到 84.29%。经过 BPSO 算法选择后的基分类器个数分别为 17、14、13、19 个，在原有基分类器个数上平均减少 34 个左右。比较不同 *k* 取值下的分类准确率，在标准随机子空间方法中，随着子空间率 *k* 值的增大，分类准确率呈现递增趋势，但经过离散二进制粒子群算法进行基分类器的选择后，分类准确率提高，但与 *k* 的取值没有相关规律。

表 5　图书评论数据分类准确率比较

k	RS	RS_BPSO
0.01	0.6867	0.8270(19)
0.02	0.7033	**0.8434**(19)
0.03	0.7633	0.8208(20)
0.05	0.785	0.8325(21)

表 5 是图书评论数据分类准确率结果。从表 5 中数据得出的结论来看与表 4 结论一致。RS_BPSO 得到的分类准确率均比标准随机子空间方法高，且高出幅度达到 5%-14%。最高分类准确率达到 84.34%。经过 BPSO 算法选择后的基分类器个数分别为 19、19、20、21 个，在原有基分类器个数上平均减少 30 个左右。随着子空间率 *k* 值的增大，标准随机子空间方法中分类准确率呈现递增趋势，但经过离散二进制粒子群算法进行基分类器的选择后，分类准确率提高，但没有呈现与 *k* 的取值相关的趋势。

表 6 是笔记本电脑评论数据分类准确率结果。RS_BPSO 得到的分类准确率均比标准随机子空间方法高，且高出幅度在 4%-7%之间。最高分类准确率达到 87.62%。经过 BPSO 算法选择后的基分类器个数分别为 24、29、28、22 个，在原有基分类器个数上平均减少 19 个左右。比较不同 *k* 取值下的分类准确率，没有发现统一的规律性。

表 6　笔记本电脑评论数据分类准确率比较

k	RS	RS_BPSO
0.01	0.7867	0.8517(24)
0.02	0.8267	**0.8762**(29)
0.03	0.8067	0.8717(28)
0.05	0.8233	0.8634(22)

从以上三个数据集在分类准确率上的对比，可以得出结论：基于 BPSO 的随机子空间方法可明显提高标准随机子空间方法分类准确率。在标准随机子空间方法中，分类准确率随着子空间率 *k* 的取值增大而提高，在基于 RS_BPSO 中影响不大。经过 BPSO 算法进行选择后的基分类器的使用个数明显减少，对提高分类系统的计算速度同时减少存储时间有很大的好处。

5.2　系统差异度结果及分析

标准 RS 和 RS_BPSO 的系统差异度分别在最终确定的基分类器对每个测试样本的输出结果上计算得到，且计算了 4 种不同的差异度度量值。表 7–表 9 分别为酒店评论数据、图书评论数据、笔记本电脑评论数据的差异度度量结果。

表 7 中，双次失败度量 *DF* 的度量中，RS_BPSO 的值均比 RS 高，说明 RS_BPSO 的差异度降低了。而在不一致度量 *dis* 中，当 *k* 取 0.01、0.02、0.03 时，RS_BPSO 的 *dis* 度量均比 RS 的 *dis* 度量高。根据不一致度量的计算原理，当 *dis* 值越大，系统差异度程度越高。而 *Q* 统计和相关系数 ρ 在计算原理上有一致的趋势。从分析数据来看，RS_BPSO 的 *Q* 统计数据和相关系数均低于 RS。根据 *Q* 统计和相关系数 ρ 的计算原理，*Q* 统计和相关系数 ρ 数值越小，则系统差异度程度越高。数据结果说明 RS_BPSO 有较高的系统差异度。

表 8 中，双次失败度量 *DF* 的度量中，RS_BPSO 与 RS 的数值差别不大，差异度没有明显不同。而在不一致度量 *dis* 中，RS_BPSO 的 *dis* 度量均比 RS 的高出 0.02 到 0.04 不等，RS_BPSO 的差异度程度更高。同样的结论在 *Q* 统计和相关系数 ρ 的数据结果中体现得更为明显。RS_BPSO 的 *Q* 统计和相关系数 ρ 均低于 RS，显然，RS_BPSO 提高了系统差异度。

表 7　酒店评论数据集成系统差异度比较

k	DF		dis		Q 统计		相关系数 ρ	
	RS	RS_BPSO	RS	RS_BPSO	RS	RS_BPSO	RS	RS_BPSO
0.01	0.3668	0.3715	0.4378	0.466	0.1507	0.0127	0.0972	0.0263
0.02	0.4396	0.4437	0.3759	0.4153	0.3794	0.1699	0.1958	0.0864
0.03	0.4677	0.4862	0.3718	0.379	0.3612	0.2837	0.179	0.136
0.05	0.5289	0.5452	0.333	0.3266	0.4448	0.4434	0.2144	0.2099

表 8　图书评论数据集成系统差异度比较

k	DF		dis		Q 统计		相关系数 ρ	
	RS	RS_BPSO	RS	RS_BPSO	RS	RS_BPSO	RS	RS_BPSO
0.01	0.321	0.3174	0.4701	0.4963	0.0667	–0.0321	0.048	–0.0099
0.02	0.3751	0.3834	0.4383	0.4585	0.1594	0.0477	0.0903	0.0351
0.03	0.4094	0.4079	0.409	0.44	0.2615	0.1071	0.1368	0.0589
0.05	0.4543	0.4576	0.3895	0.4115	0.2935	0.1663	0.1448	0.079

表 9 中，可以得到和表 8 同样的结论。双次失败度量 *DF* 的度量中，RS_BPSO 比 RS 差别不大，差异度没有明显差异。而在不一致度量 *dis* 中，RS_ BPSO 的 *dis* 度量均比 RS 的 *dis* 度量高，RS_BPSO 的差异度程度高。RS_BPSO 的 *Q* 统计和相关系数 ρ 均低于 RS，RS_BPSO 的差异度显然高于标准特征子空间方法。

表 9　笔记本电脑评论文本集成系统差异度比较

k	DF		dis		Q 统计		相关系数 ρ	
	RS	RS_BPSO	RS	RS_BPSO	RS	RS_BPSO	RS	RS_BPSO
0.01	0.3284	0.3271	0.4722	0.4986	0.0422	–0.0616	0.0399	–0.021
0.02	0.3753	0.3796	0.4559	0.4629	0.0482	0.0233	0.061	0.0265
0.03	0.4114	0.4073	0.428	0.441	0.1462	0.077	0.0875	0.057
0.05	0.4731	0.4764	0.3879	0.3909	0.2504	0.2225	0.1276	0.1146

为了更加直观地分析不同 *k* 取值下标准随机子空间和离散二进制粒子群随机子空间方法的差异度，图 3 分别为酒店评论数据差异度对比图、图书评论数据差异度对比图、笔记本电脑评论数据差异度对比图。

从图 3 可以看出，*DF*、*Q* 统计、相关系数 ρ 呈现上升趋势，而 *dis* 呈现下降趋势。4 种差异度度量得出一致结论，随着 *k* 取值的增加，无论在标准随机子空间方法还是在 RS_BPSO 中，差异度都呈现递减的趋势。而在 RS 与 RS_BPSO 的两两比较中，双次失败度量 *DF*、*Q* 统计和相关系数 ρ 在 RS_BPSO 中得到的曲线均在 RS 的曲线之上，不一致度量 *dis* 则相反。根据其计算理论，得出一致结论，RS_BPSO 的差异度高于标准随机子空间方法。图书评论数据和笔记本电脑评论数据都得出与酒店评论数据一样的结论。

从以上分析得出结论：标准随机子空间方法和 RS_BPSO 的系统差异度都随着 *k* 值的增加而降低，RS_BPSO 的差异度明显高于标准随机子空间方法。进而说明 RS_BPSO 的泛化能力强于标准随机子空间方法。

而与分类准确率结果比较，在标准随机子空间方法中，随着 *k* 值的增加，分类准确率上升。差异度度量与分类准确率呈现矛盾的趋势。这与 Chandra 等提出的想法一致：系统分类器准确率和差异度之间存在一个权衡(Trade-Off)[17]。通过负相关集成学习理论提出多目标优化集成学习，即将准确率和差异度作为两个优化目标，用来作为达到准确率与差异度权衡的一个策略。而另一部分文献改进了现有差异度度量方法，将分类器准确率和差异度组合起来构成复合差异度函数[18-19]。构造分类准确率与系统差异度集合的函数作为智能算法适应度函数也将取得很多成果。

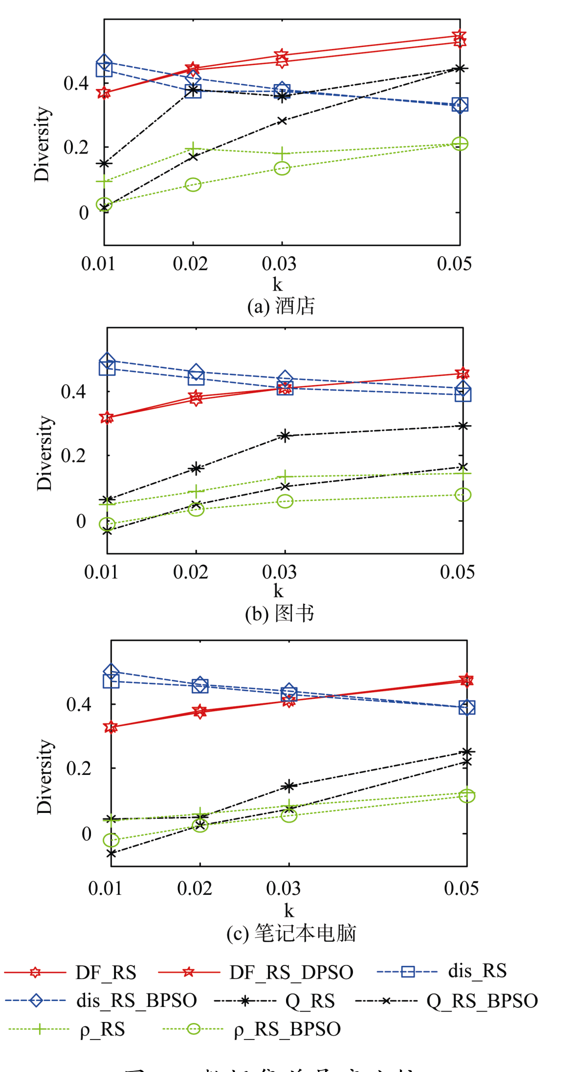


图 3　数据集差异度比较

鉴于 BPSO 算法在提高分类准确率方面对 k 值的敏感度较低，而在系统差异度即系统泛化能力上随着 k 值增加而降低，则最佳 k 值应选用较小的 k 值。与此同时，考虑到 k 值的选取是基分类器训练样本文本表示特征维数的确定，关系到基分类器训练样本特征表示，基分类器准确率的好坏依赖于样本特征表示。所以，k 值的选取应在保证基分类器一定准确率的情况下选用较小值。

5.3　BPSO 算法优化过程分析

(1) BPSO 收敛性分析

BPSO 收敛性是指随着迭代次数的增加，算法的结果与真实结果的误差越来越小，且趋近于一个固定值。收敛与发散对应，发散是指无论迭代次数多大，收敛曲线无法趋于定值。

为了更好地分析 BPSO 算法在集成学习随机子空间方法基分类器的优化性能，对 BPSO 算法的收敛性进行分析。对 BPSO 算法的 100 次迭代得到的最好适应值结果和平均适应值结果绘制了曲线。图 4–图 6 分别为酒店评论数据、图书评论数据、笔记本电脑评论数据的适应值对迭代次数的变化图。

从图 4 可以看出，酒店评论数据文本情感分类最好适应值随着迭代次数的增加而不断地增长，说明分类误差越来越小，逐渐趋近于 0。而平均适应值虽然局部出现震荡但整体呈现向分类准确率最大值逼近的趋势，说明 BPSO 算法在随机子空间基分类器的优化选择中收敛性很好。

从图 5 可以看出，图书评论数据文本情感分类最好适应值随着迭代次数的增加而不断地增长，而平均适应值虽然局部出现震荡但在整体趋势上依然呈现上升趋势。

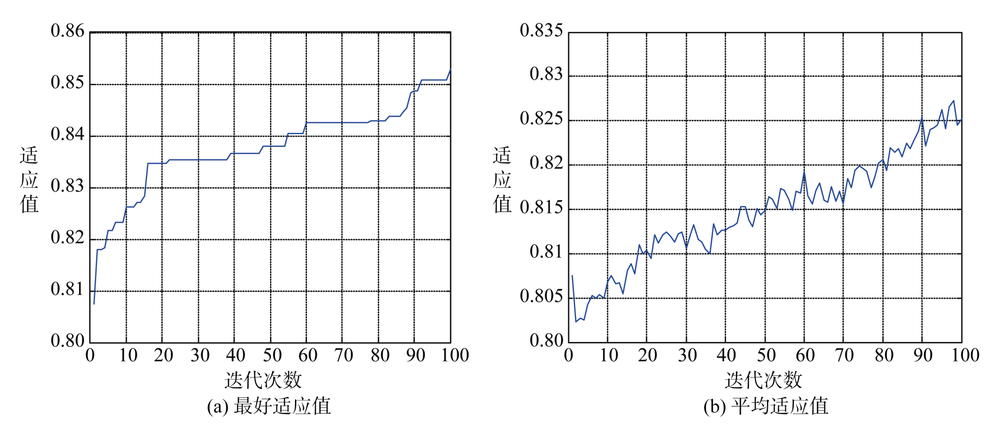


图 4　酒店评论数据适应值对迭代次数的变化

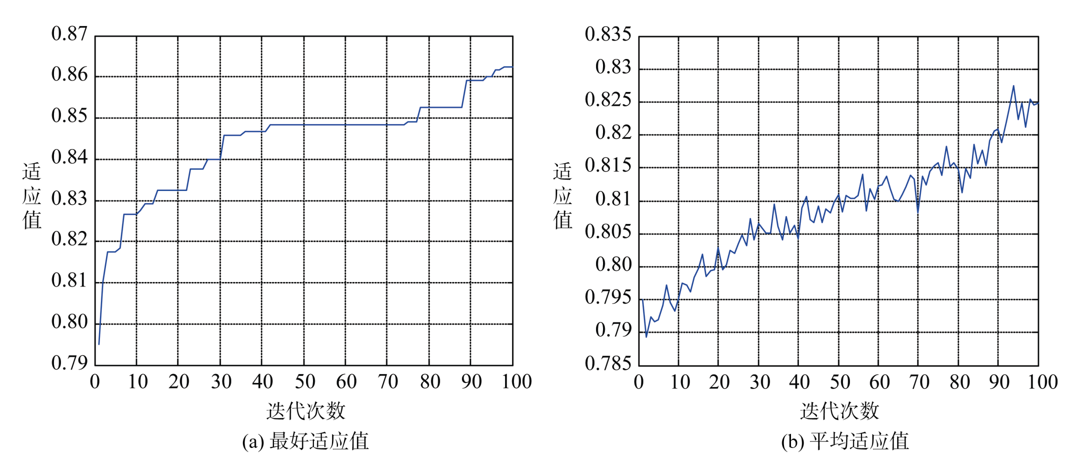


图 5　图书评论数据适应值对迭代次数的变化

如图 6 所示，笔记本电脑评论数据文本情感分类适应值函数随着迭代次数的增加而呈现不断上升趋势，平均适应度函数值整体呈现上升趋势。

从以上分析可以得出，BPSO 算法在以文本情感分类准确率为目标适应值函数时的收敛性很好，可以提高原有分类准确率。

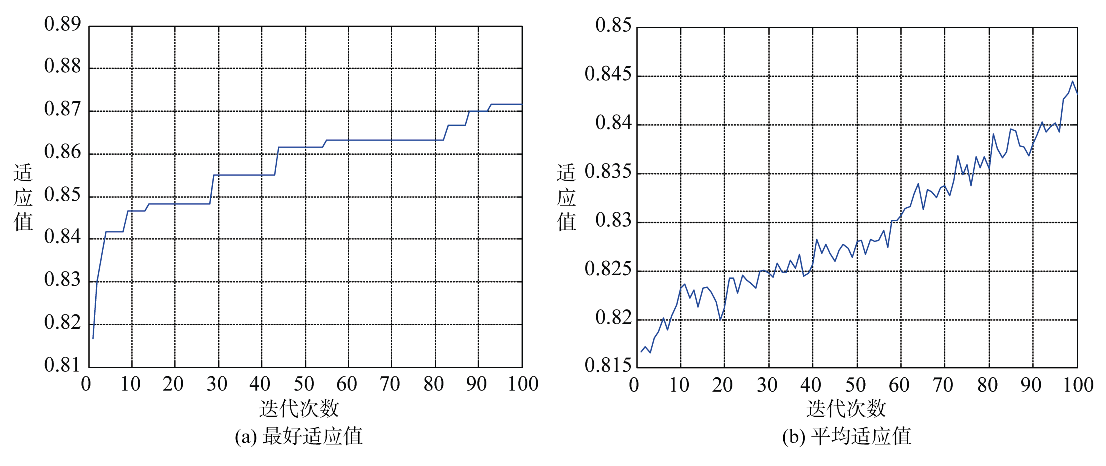


图 6　笔记本电脑评论数据适应值对迭代次数的变化

(2) 粒子个数对分类结果的影响

以上均是粒子群个数为 50 的分类结果，为了探索不同的粒子数对文本情感分类的影响，分别采用粒子群个数为 10、20、30、40、50、60、70、80、90、100 进行实验。图 7 是以笔记本电脑评论数据集为例给出的分类准确率随粒子数目变化趋势图。粒子数目取 30 时，分类准确率达到最高。在粒子群算法优化过程中，粒子数量对集成系统的效果有一定的影响。一般情况下，集成系统的性能会随着粒子数量的增大而有所提高。粒子数目越多，其搜索到全局最优区域的速度更快，相反如果粒子数越少，搜索时间可能更长。同时，粒子群过少，陷入局部最优的可能性则越大。但是，粒子数目过多时，由于每个粒子都要重复计算适应度值并调整自己的方向和速度，因此时间代价也会很高。可以看出，在文本情感分类问题中，30 个粒子达到最好分类准确率，而随着粒子数目的增加，分类准确率有轻微的下降趋势。而在酒店评论数据、图书评论数据上也得到相同的结论。因此，在本实验数据的规模上，粒子数目选择 30 左右比较合适。

通过以上对标准随机子空间和 BPSO 随机子空间两种方法在分类准确率上的比较和分析、对集成系统差异度的影响和分析以及 BPSO 算法的优化过程进行分析，可以得出结论，RS_BPSO 分类准确率在标准随机子空间方法上提高了 3%-15%，系统差异度上也得到明显提高。RS_BPSO 的使用减少了分类预测系统的基分类器个数，对提高分类系统的计算速度同时减少

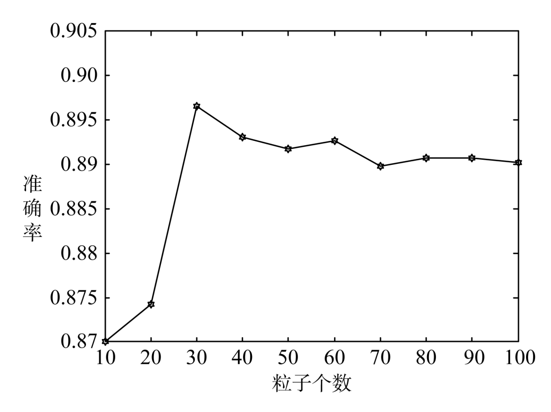


图 7 笔记本电脑数据集上粒子数目与分类准确率趋势图(k=0.01)

存储空间有很大的帮助。同时发现，子空间率的取值本来是标准随机子空间方法中准确率和差异度难以调和的参数，但经过 BPSO 算法后，子空间率不再影响分类准确率，而只与系统差异度有关。根据系统差异度对子空间率的变化趋势，子空间率的选取规则应为在给定数值范围内选择较小值。对 BPSO 在基分类器个数的优化选择过程中的收敛性分析，说明 BPSO 算法可以很好地应用于随机子空间方法中。RS_BPSO 在标准随机子空间方法上提高了分类准确率。

6 结 语

文本情感分类技术在电子商务、电子政务、信息预测等领域有重要的应用价值。针对中文文本情感表达多样性和隐晦性的特点，以依存句法解析特征表示为基础，提出基于BPSO的随机子空间集成分类方法。随机子空间方法以划分特征的方式形成个体分类器的训练数据，一定程度上降低了训练模型中数据输入的维度。BPSO 作为选择性机制，既提高了集成系统的分类准确率也保证了系统的泛化能力。同时，对比研究子空间率对标准随机子空间和 BPSO 随机子空间方法的分类准确率和系统差异度的影响，得出了 BPSO 随机子空间方法中子空间率选值的一般规律。

在未来研究中，一方面会搜集更多的数据集包括英文数据集进一步验证本文的结论；另一方面针对本文已得到的结论，构造更适合文本情感分类的类模型。

参考文献：

[1] Agarwal B, Mittal N. Machine Learning Approach for Sentiment Analysis [A]// Prominent Feature Extraction for Sentiment Analysis[M]. Springer, International Publishing, 2016: 21-45.

[2] Vinodhini G, Chandrasekaran R. Sentiment Analysis and Opinion Mining: A Survey[J]. International Journal of Advanced Research in Computer Science and Software Engineering, 2012, 2(6): 282-292.

[3] Liu B, Zhang L. A Survey of Opinion Mining and Sentiment Analysis [A].// Mining Text Data[M]. Springer US, 2012.

[4] 张庆庆，刘西林. 基于依存句法关系的文本情感分类研究[J]. 计算机工程与应用，2015, 51(22): 28-32. (Zhang Qingqing, Liu Xilin. Sentiment Analysis Based on Dependency Sytactic Relation[J]. Computer Engineering and Applications, 2015, 51(22): 28-32.)

[5] Wang G, Sun J, Ma J, et al. Sentiment Classification: The Contribution of Ensemble Learning[J]. Decision Support Systems, 2014, 57(1): 77-93.

[6] Wang G, Zhang Z, Sun J, et al. POS-RS: A Random Subspace Method for Sentiment Classification Based on Part-of-Speech Analysis[J]. Information Processing & Management, 2015, 51(4): 458-479.

[7] Dasarathy B V, Sheela B V. A Composite Classifier System Design: Concepts and Methodology[J]. Proceedings of the IEEE, 1979, 67(5): 708-713.

[8] Polikar R. Ensemble Based Systems in Decision Making[J]. IEEE Circuits and Systems Magazine, 2006, 6(3): 21-45.

[9] Dietterich T G. Ensemble Methods in Machine Leanring[C]// Proceedings of the 1st International Workshop on Multiple Classifier Systems.2000.

[10] Ho T K. The Random Subspace Method for Constructing Decision Forests[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(8): 832-844.

[11] 孙博，王建东，陈海燕，等. 集成学习中的多样性度量[J]. 控制与决策, 2014, 29(3): 385-395. (Sun Bo, Wang Jiandong, Chen Haiyan, et al. Diversity Measures in Ensemble Learning [J]. Control and Decision, 2014, 29(3): 385-395.

[12] Zhou Z H, Wu J X, Jiang Y, et al. Genetic Algorithm Based Selective Neural Network Ensemble[C]// Proceedings of the 17th International Joint Conference on Artificial Intelligence. 2001.

[13] Tama B A, Rhee K H. A Combination of PSO-Based Feature Selection and Tree-Based Classifiers Ensemble for Intrusion Detection Systems [A].// Advances in Computer Science and Ubiquitous Computing[M]. Singapore: Springer, 2015.

[14] Hedeshi N G, Abadeh M S. Coronary Artery Disease

Detection Using a Fuzzy-boosting PSO Approach [J]. Computational Intelligence and Neuroscience, 2014, 2014: Article No. 783734. http://dx.doi.org/10.1155/2014/783734.

[15] Tsai C Y, Chen C J. A PSO-AB Classifier for Solving Sequence Classification Problems [J]. Applied Soft Computing, 2015, 27: 11-27.

[16] Kennedy J, Eberhart R C. A Discrete Binary Version of the Particle Swarm Algorithm[C]//Proceedings of the 1997 Conference on Systems, Man, and Cybernetics. 1997: 4104-4108.

[17] Chandra A, Chen H, Yao X. Trade-off Between Diversity and Accuracy in Ensemble Generation [A]// Multi-objective Machine Learning[M]. Springer Berlin Heidelberg, 2006.

[18] Ko A H R, Sabourin R, De Souza Britt Jr A. Combining Diversity and Classification Accuracy for Ensemble Selection in Random Subspaces[C]//Proceedings of the International Joint Conference on Neural Networks.2006.

[19] Ko A H R, Sabourin R, De Souza Britto Jr A. Compound Diversity Functions for Ensemble Selection[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2009, 23(4): 659-686.

作者贡献声明:

张庆庆: 提出研究思路, 设计研究方案, 完成实验;
刘西林: 论文起草及最终版本修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: suiyue2959@163.com。
[1] 张庆庆. hotel4000.xls.酒店评论数据.
[2] 张庆庆. book2000.xls. 图书评论数据.
[3] 张庆庆. notebook1000.xls. 笔记本电脑评论数据.
[4] 张庆庆. useRSTraining.m. 随机子空间训练程序.
[5] 张庆庆. GetDependencyAndDotlines.java. 三元组依存关系特征抽取程序.
[5] 张庆庆. main.m. BPSO 训练程序.
[6] 张庆庆. batchRSPSO.m. RS_BPSO 训练程序.

收稿日期: 2017-03-28
收修改稿日期: 2017-04-24

Classifying Sentiments Based on BPSO Random Subspace

Zhang Qingqing[1,2]　Liu Xilin[2]
[1](School of Management, Xi'an Polytechnic University, Xi'an 710048, China)
[2](School of Management, Northwestern Polytechnical University, Xi'an 710129, China)

Abstract: [**Objective**] This paper aims to solve the issue of representing high dimensional features in Chinese sentiment analysis, with the help of RS_BPSO, a selective ensemble algorithm. [**Methods**] First, we developed the framework and algorithm of the proposed RS_BPSO model based on the theory of Random Subspace and Binary Particle Optimization. Then, we transformed the Chinese review corpus into structured feature vectors and examined the new model. [**Results**] We found that the diversity and accuracy of the RS_BPSO model better than the standard RS model. [**Limitations**] We did not run the proposed model with corpus in foreign languages. [**Conclusions**] The RS_BPSO model could be an effective method to classify Chinese sentiments.

Keywords: Random Subspace　BPSO　Text Sentiment Classification　Subspace Rate